Published online 17 November 2012 *Nucleic Acids Research, 2013, Vol. 41, Database issue* ***D43–D47***
doi:10.1093/nar/gks1068

Update on activities at the Universal Protein Resource (UniProt) in 2013

The UniProt Consortium[1,2,3,4,*]

[1]The EMBL Outstation, The European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK, [2]SIB Swiss Institute of Bioinformatics, Centre Medical Universitaire, 1 rue Michel Servet, 1211 Geneva 4, Switzerland, [3]Protein Information Resource, Georgetown University Medical Center, 3300 Whitehaven Street North West, Suite 1200, Washington, DC 20007 and [4]Protein Information Resource, University of Delaware, 15 Innovation Way, Suite 205, Newark, DE 19711, USA

Received September 26, 2012; Revised and Accepted October 11, 2012

ABSTRACT

The mission of the Universal Protein Resource (UniProt) (http://www.uniprot.org) is to support biological research by providing a freely accessible, stable, comprehensive, fully classified, richly and accurately annotated protein sequence knowledgebase. It integrates, interprets and standardizes data from numerous resources to achieve the most comprehensive catalogue of protein sequences and functional annotation. UniProt comprises four major components, each optimized for different uses, the UniProt Archive, the UniProt Knowledgebase, the UniProt Reference Clusters and the UniProt Metagenomic and Environmental Sequence Database. UniProt is produced by the UniProt Consortium, which consists of groups from the European Bioinformatics Institute (EBI), the SIB Swiss Institute of Bioinformatics (SIB) and the Protein Information Resource (PIR). UniProt is updated and distributed every 4 weeks and can be accessed online for searches or downloads.

INTRODUCTION

The UniProt's goal is to provide the most comprehensive resource for protein sequence and functional annotation. The four UniProt databases are optimized for different uses as follows: the UniProt Knowledgebase (UniProtKB) is an expertly curated database; the UniProt Archive (UniParc) (1) is a comprehensive sequence repository, reflecting the history of all protein sequences not only in the UniProtKB but also in all source databases; the UniProt Reference Clusters (UniRef), which merge closely related sequences based on sequence identity to facilitate sequence similarity searches (2) and the UniProt Metagenomic and Environmental Sequence (UniMES) database, which was created to cater for the developing area of metagenomics. The aim of this article is to provide a status report on UniProt activities and some of our plans for the near future that will enable us to successfully continue to play a critical role in bioinformatics discovery in the genomic and proteomic era.

NEW AND ONGOING DEVELOPMENTS

UniProtKB reorganization

As the cost of sequencing continues to fall, the number of organisms with complete proteomes in UniProtKB is increasing. It is also becoming more and more common in the scientific community for many groups to sequence the complete proteomes of the same organism or multiple strains of an organism. This means that users are presented with an increasingly large data set, which can be difficult to navigate and are largely redundant in biological knowledge. In response, the UniProt Consortium is developing a concept to provide a set of sequences from selected species based on UniProtKB manually reviewed entries, the Reference Proteomes and the Representative proteomes (3,4). We re-evaluated our manual annotation priorities, and re-defined our organism focus list. For more information, please see http://www.uniprot.org/program. Curators continue to define complete proteomes and reference proteomes as they become available. To ensure comprehensiveness, several changes were required in the UniProt import pipeline. Historically, the great majority of UniProt sequences are based on translations of genome sequence submissions to the International Nucleotide Sequence Database Consortium (INSDC) (5). Our longstanding collaboration has been deepened to include the joint definition of complete genomes and the grouping together of all the genome submissions

*To whom correspondence should be addressed. Tel: +44 1223 494435; Fax: +44 1223 494468; Email: apweiler@ebi.ac.uk

(e.g. individual chromosomes, organelles) for an organism that originate from the same sequencing project under one unique set accession. In addition, we have extended the import pipeline to include Ensembl (6) and Ensembl Genomes (7) sequences. This was to ensure comprehensiveness, as the full and/or up-to-date annotation of genomes is sometimes not submitted to the INSDC, for example, *Apis mellifera* (http://metazoa.ensembl.org/Apis_mellifera/Info/Index). The Ensembl sequences are mapped to their UniProtKB counterparts under stringent conditions, requiring 100% identity for 100% of the length of the two sequences. Ensembl sequences that are absent from UniProtKB are imported into UniProtKB/TrEMBL. The UniProtKB entries provide a cross-reference back to the appropriate Ensembl record(s) where available, enabling an easy transition to the genomic view. The one exception to this approach is for the *Homo sapiens* complete proteome, where there are some cross-references to Ensembl in the UniProtKB/Swiss-Prot entries that do not follow the aforementioned criteria. This is because of the fact that there are different evidence and sources for the sequence in the two resources. The cross-reference mapping is, however, enhanced with the usage of HUGO Gene Nomenclature Committee (HGNC) (http://www.genenames.org) identifiers. Of the 20 224 UniProtKB/Swiss-Prot entries, 18 696 entries have at least one sequence that has 100% identity for 100% of the length of an Ensembl transcript. The UniProt curators and the Ensembl curators and gene builders are progressively working through the rest of the differences, correcting them where appropriate and documenting agree-to-disagree decisions. This is part of the Consensus CDS (CCDS) project, which is a collaborative effort to identify a core set of human and mouse protein coding regions that are consistently annotated and of high quality (http://www.ncbi.nlm.nih.gov/CCDS/). The long-term goal is to support convergence towards a standard set of gene annotations. UniProt has also extended the pipeline to import RefSeq (8) sequences, and we are currently evaluating how to combine this data with the existing UniProtKB and Ensembl data. All of these developments have had the side benefit of establishing a close and mutually beneficial collaboration with the Ensembl and RefSeq groups. We import their sequences while they import our annotations into their records (in particular the protein nomenclature and sequence feature annotations), and their prediction pipelines learn from our manually reviewed and experimentally proved sequences. There is a consensus that we should all provide the same (in sequence and annotation) complete proteomes and to collaborate on the definition of Reference proteomes. Another outcome of this collaboration is the ongoing development of genome annotation standards (including protein nomenclature), and the promotion of these standards by the sequencing community (9).

UniProt biocuration

UniProt's central focus is the annotation—both manual and automatic—of the UniProt Knowledgebase.

Manual curation challenge

Historically, the sequences from the same gene (and more than one when the resulting protein sequences were 100% identical) from the same organism were merged into one UniProtKB/Swiss-Prot entry. Discrepancies between sequence reports were identified, and the underlying causes, such as alternative splicing, natural variations, frameshifts and so forth, were annotated. Journal articles provided the main source of experimental knowledge, with the full text of each article being read and the information extracted. The aim of this approach was to provide a central hub of information for each protein, but it also meant that many UniProtKB/Swiss-Prot entries contain sequences and annotations from many strains. In the era of complete genomes and proteomes at the strain level for so many organisms, UniProt has modified this policy. We are now providing entries that contain the protein products from a particular gene from a particular species + strain with the experimental literature being annotated to that species + strain and propagated as appropriate to other species and strains, ideally through the UniRule pipeline (see later in the text). This has the advantage of providing a gold standard experimental set in UniProtKB/Swiss-Prot and automatically propagating appropriate annotation to the ever increasing number of complete proteomes for which there is no experimental data in UniProtKB/TrEMBL.

Automatic annotation approaches

UniProt has developed two complementary systems to automatically annotate the protein sequences in UniProtKB/TrEMBL. The first system, UniRule, which incorporates the HAMAP (10), RuleBase (11) and PIR Rule (12,13) systems, consists of annotation rules created and monitored by experienced curators. Each annotation rule specifies a number of annotations and conditions which must be satisfied for that annotation to be applied. These conditions may include family membership [as indicated by a match to a family defined by InterPro (14)], taxonomic constraints and the presence of particular sequence features. Rules are created by curators based on information from experimentally characterized template entries, and their predictions evaluated against the content of manually annotated UniProtKB/Swiss-Prot entries, which serve as the gold standard. With each UniProt release, the monitoring system sends those rules that are inconsistent with UniProtKB/Swiss-Prot annotation to curators for review. This ensures that only high-quality predictions are added and prevents propagation of potentially erroneous data. The second system, the Statistical Automatic Annotation System [SAAS, previously named Spearmint (15)] supplements the labour-intensive-UniRule system and generates automatic rules for functional annotation from UniProtKB/Swiss-Prot entries using the C4.5 decision-tree algorithm. This algorithm uses entropy gain to find the most concise rule for an annotation based on the criteria of sequence length, InterPro-group membership and taxonomy. Generating rules 'on the fly' ensure their evolution along with the UniProtKB with little or no manual intervention while providing seed rules for exploitation in the UniRule

system. This combined approach produces annotation for 34% of UniProtKB/TrEMBL entries at the current time. All predictions are refreshed with each UniProtKB release to ensure the latest state-of-knowledge predictions.

Gene Ontology annotation

UniProt continues to be a major provider of Gene Ontology (GO) annotations to the GO Consortium (16). UniProt curators are actively involved in curating UniProtKB entries with GO terms, providing both high-quality manual GO annotations in addition to their contributions to electronic GO annotation pipelines. Manual GO annotations are made during the UniProt literature curation process, and, at the time of writing, almost 214 000 annotations have been manually assigned to >37 000 proteins by UniProt curators. The curators also supply information to entries that is subsequently used in electronic GO annotation pipelines, such as UniProt keywords2GO, UniProt subcellular location2GO and InterPro2GO. A new automatic pipeline, UniPathway2GO [a collaboration between UniProt, INRIA (Rhone-Alpes) and Laboratoire d'Ecologie Alpine (Grenoble) (17)], was initiated in May 2012 that provides GO annotations describing the metabolic pathways that proteins are involved in. Altogether, the UniProt supplied automatic annotation pipelines provide 42.5 million annotations to >14 million proteins. UniProt also incorporates annotations from other GO Consortium members and affiliates and displays these annotations in the relevant UniProt entries. Currently, the UniProt-GO annotation project provides GO annotations for 65% of UniProt entries.

Highlighting the UniProt website

As a result of recent usability testing with the UniProt user community, we would like to highlight the following features on the UniProt website (http://www.uniprot.org), which is the main access point for the data available in the UniProt databases and the tools to explore it. The tabbed bar on the top of each page includes multiple tools, such as free text 'Search', 'BLAST' sequence similarity search, 'Align' for multiple sequence alignment, 'Retrieve' for batch downloads and 'ID mapping'. ID mapping is a tool to convert UniProt identifiers to corresponding identifiers from a number of other databases available in a dropdown list or vice versa. There is also functionality available to help users personalize their experience with the website. For example, the search results page contains the 'Customize' button above the results table to help modify the table. This allows removal or addition of data to the results table from a vast selection of available columns, such as Gene Ontology, Cross-references, Sequence features and so forth, to help users find their proteins of interest. Users can then click on checkboxes at the left of the results table to add their proteins of interest to a selection cart that appears at the bottom of the page. The cart provides tools to help analyse or download the selected entries and saves selections across searches. The protein entry page contains the 'Customize order' button on the grey navigation tool bar that allows users to reorder sections within the entry.

DATABASE ACCESS AND FEEDBACK

The http://www.uniprot.org website (18) is the primary access point to our data and documentation and offers tools, such as full text and field-based text search, sequence similarity search, multiple sequence alignment, batch retrieval and database identifier mapping. The home page features a site tour as a quick introduction for novice users. The full text search allows quick and easy searching without previous knowledge of our data or search syntax. The results are sorted by relevance, and search suggestions are provided, where possible, to help filter searches that yield too many or no results. More complex queries can be built with the field-based text search, either iteratively with a query builder or by entering them manually in the query field, which can be faster and more powerful (http://www.uniprot.org/help/text-search). Searching with ontology terms is assisted by auto-completion, and search results can be browsed by ontologies. The display of the result sets, as well as database entries, is configurable; columns can be added to or removed from the result table to see more functional annotation than is available in the default display. Sequence similarity search results can be filtered by taxonomy to obtain a quick overview of the taxonomic distribution of the results, and the sequence annotations of the matched entries can be projected onto the sequence alignments to see at a glance whether important positions are conserved. The site has a simple and consistent URL scheme that allows the bookmarking of all searches to repeat them at a later time. All result sets can be downloaded to offer users the possibility to retrieve customized data sets. However, large downloads are given low priority to ensure that they do not interfere with interactive queries, and they can, therefore, be slow compared with downloads from the UniProt FTP server. We, therefore, recommend downloading complete data sets from ftp.uniprot.org/pub/databases. The website offers various download formats (e.g. plain text, extensible mark-up language, RDF, FASTA, GFF), which depend on the chosen data set. The tab-delimited and Excel formats can be customized by selecting the desired columns in the graphical view of the result table. All data are also available in RDF (http://www.w3.org/RDF/), a W3C standard for publishing data on the Semantic Web. Both data and search results can also be accessed programmatically, either through simple HTTP (REST) requests (http://www.uniprot.org/faq/28) or our Java API (UniProtJAPI) (19).

Although the UniProt website provides a query interface for all UniProt data, some users also require facilities to search across related data in different databases. We have, therefore, set-up a BioMart (20) (http://www.biomart.org) instance at http://www.ebi.ac.uk/uniprot/biomart/martview that allows complex queries between UniProt and other data resources, such as PRIDE (21), Ensembl and InterPro. To offer users even more

flexibility, we are going to provide a SPARQL Protocol and RDF Query Language (SPARQL) (http://www.w3.org/TR/rdf-sparql-query/) end-point for all our data that can be linked with any remote data resource that has a SPARQL end-point, using SPARQL 1.1's federated query capabilities. This new service is available for beta testing at http://beta.sparql.uniprot.org/.

Your feedback is extremely valuable to help us improve our databases and services in terms of accuracy and usability. Please contact us if you have questions or suggestions through http://www.uniprot.org/contact or email us directly at help@uniprot.org. You can submit new data or updates at http://www.uniprot.org/help/submissions. Extensive documentation on how to best use our resource is available at http://www.uniprot.org/help/. UniProt is freely available for both commercial and non-commercial use. Please see http://www.uniprot.org/help/license for details. New releases are published every 4 weeks except for UniMES, which is updated only when the underlying source data are updated. Release statistics are available at http://www.uniprot.org.

ACKNOWLEDGEMENTS

UniProt has been prepared by Rolf Apweiler, Maria Jesus Martin, Claire O'Donovan, Michele Magrane, Yasmin Alam-Faruque, Emanuela Alpi, Ricardo Antunes, Joanna Arganiska, Elisabet Barrera Casanova, Benoit Bely, Mark Bingley, Carlos Bonilla, Ramona Britto, Borisas Bursteinas, Wei Mun Chan, Gayatri Chavali, Elena Cibrian-Uhalte, Alan Da Silva, Maurizio De Giorgi, Emily Dimmer, Francesco Fazzini, Paul Gane, Alexander Fedotov, Leyla Garcia Castro, Penelope Garmiri, Emma Hatton-Ellis, Reija Hieta, Rachael Huntley, Julius Jacobsen, Rachel Jones, Duncan Legge, Wudong Liu, Jie Luo, Alistair MacDougall, Prudence Mutowo, Andrew Nightingale, Sandra Orchard, Samuel Patient, Klemens Pichler, Diego Poggioli, Sangya Pundir, Luis Pureza, Guoying Qi, Steven Rosanoff, Tony Sawford, Harminder Sehra, Edward Turner, Vladimir Volynkin, Tony Wardell, Xavier Watkins, Hermann Zellner, Matt Corbett, Mike Donnelly, Pieter van Rensburg, Mickael Goujon, Hamish McWilliam and Rodrigo Lopez at the European Bioinformatics Institute (EBI); Ioannis Xenarios, Lydie Bougueleret, Alan Bridge, Sylvain Poux, Nicole Redaschi, Andrea Auchincloss, Kristian Axelsen, Parit Bansal, Delphine Baratin, Pierre-Alain Binz, Marie-Claude Blatter, Brigitte Boeckmann, Jerven Bolleman, Emmanuel Boutet, Lionel Breuza, Edouard de Castro, Lorenzo Cerutti, Elisabeth Coudert, Beatrice Cuche, Mikael Doche, Dolnide Dornevil, Severine Duvaud, Anne Estreicher, Livia Famiglietti, Marc Feuermann, Elisabeth Gasteiger, Sebastien Gehant, Vivienne Gerritsen, Arnaud Gos, Nadine Gruaz-Gumowski, Ursula Hinz, Chantal Hulo, Janet James, Florence Jungo, Guillaume Keller, Vicente Lara, Philippe Lemercier, Jocelyne Lew, Damien Lieberherr, Xavier Martin, Patrick Masson, Anne Morgat, Teresa Neto, Salvo Paesano, Ivo Pedruzzi, Sandrine Pilbout, Monica Pozzato, Manuela Pruess, Catherine Rivoire, Bernd Roechert, Michel Schneider, Christian Sigrist, Karin Sonesson, Sylvie Staehli, Andre Stutz, Shyamala Sundaram, Michael Tognolli, Laure Verbregue, Anne-Lise Veuthey and Mohamed Zerara at the SIB Swiss Institute of Bioinformatics (SIB); Cathy H. Wu, Cecilia N. Arighi, Leslie Arminski, Chuming Chen, Yongxing Chen, Hongzhan Huang, Abhishek Kukreja, Kati Laiho, Peter McGarvey, Darren A. Natale, Thanemozhi G. Natarajan, Natalia V. Roberts, Baris E. Suzek, C. R. Vinayaka, Qinghua Wang, Yuqi Wang, Lai-Su Yeh, Meher Shruti Yerramalla and Jian Zhang at the Protein Information Resource (PIR).

FUNDING

National Institutes of Health (NIH) [1U41HG006104-03]; NIH GO [2P41HG02273-07, 5R01GM080646-07, 3R01GM080646-07S1, 5G08LM010720-03 and 8P20GM103446-12]; British Heart Foundation [SP/07/007/23671]; Swiss Federal Government through the Federal Office of Education and Science; EC [SLING (226073), GEN2PHEN (200754) and MICROME (222886)]; National Science Foundation (NSF) [DBI-1062520]. Funding for open access charge: NIH [1U41HG006104-03].

Conflict of interest statement. None declared.

REFERENCES

1. Leinonen,R., Diez,F.G., Binns,D., Fleischmann,W., Lopez,R. and Apweiler,R. (2009) UniProt archive. *Bioinformatics*, **20**, 3236–3237.
2. Suzek,B.E., Huang,H., McGarvey,P., Mazumder,R. and Wu,C.H. (2007) UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics*, **23**, 1282–1288.
3. The UniProt Consortium. (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **40**, D71–D75.
4. Chen,C., Natale,D.A., Finn,R.D., Huang,H., Zhang,J., Wu,C.H. and Mazumder,R. (2011) Representative proteomes: a stable, scalable and unbiased proteome set for sequence analysis and functional annotation. *PLoS One*, **6**, e18910.
5. Karsch-Mizrachi,I., Nakamura,Y., Cochrane,G.; International Nucleotide Sequence Database Collaboration. (2012) The International Nucleotide Sequence Database Collaboration. *Nucleic Acids Res.*, **40**, D33–D37.
6. Flicek,P., Amode,M.R., Barrell,D., Beal,K., Brent,S., Carvalho-Silva,D., Clapham,P., Coates,P., Fairley,S., Fitzgerald,S. *et al.* (2012) Ensembl 2012. *Nucleic Acids Res.*, **40**, D84–D90.
7. Kersey,P.J., Staines,D.M., Lawson,D., Kulesha,E., Derwent,P., Humphrey,J.C., Hughes,D.S.T., Keenan,S., Kerhornou,A., Koscielny,G. *et al.* (2011) Ensembl Genomes: an integrative resource for genome-scale data from non-vertebrate species. *Nucleic Acids Res.*, **40**, D91–D97.
8. Pruitt,K.D., Tatusova,T., Brown,G.R. and Maglott,D.R. (2012) NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res.*, **40**, D130–D135.
9. Klimke,W., O'Donovan,C., White,O., Brister,J.D., Clark,K., Fedorov,B., Mizrachi,I., Pruitt,K.D. and Tatusova,T. (2011) Solving the Problem: Genome Annotation Standards before the Data Deluge. *Stand. Genomic Sci.*, **5**, 168–193.
10. Lima,T., Auchincloss,A.H., Coudert,E., Keller,G., Michoud,K., Rivoire,C., Buillard,V., de Castro,R., Lachaize,C., Baratin,D. *et al.* (2009) HAMAP: a database of completely sequenced microbial proteome sets and manually curated microbial protein families in UniProtKB/Swiss-Prot. *Nucleic Acids Res.*, **37**, D471–D478.

11. Fleischmann,W., Moller,S., Gateau,A. and Apweiler,R. (1999) A novel method for automatic functional annotation of proteins. *Bioinformatics*, **15**, 228–233.
12. Natale,D.A., Vinayaka,C.R. and Wu,C.H. (2004) Large-scale, classification-driven, rule-based functional annotation of proteins. In: Subramaniam,S. (ed.), *Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics, Bioinformatics Volume*. John Wiley & Sons, NY.
13. Vasudevan,S., Vinayaka,C.R., Natale,D.A., Huang,H., Kahsay,R.Y. and Wu,C.H. (2011) Structure-guided rule-based annotation of protein functional sites in UniProt knowledgebase. *Methods Mol. Biol.*, **694**, 91–105.
14. Hunter,S., Jones,P., Mitchell,A., Apweiler,R., Attwood,T.K., Bateman,A., Bernard,T., Binns,D., Bork,P., Burge,S. *et al.* (2012) InterPro in 2011: new developments in the family and domain predication database. *Nucleic Acids Res.*, **40**, D306–D312.
15. Kretschmann,E., Fleischmann,W. and Apweiler,R. (2001) Automatic rule generation for protein annotation with the C4.5 data mining algorithm applied on Swiss-Prot. *Bioinformatics*, **17**, 920–926.
16. Dimmer,E.C., Huntley,R.P., Alam-Faruque,Y., Sawford,T., O'Donovan,C., Martin,M.J., Bely,B., Browne,P., Mun Chan,W., Eberhardt,R. *et al.* (2012) The UniProt-GO Annotation database in 2001. *Nucleic Acids Res.*, **40**, D559–D564.
17. Morgat,A., Coissac,E., Coudert,E., Axelsen,K.B., Keller,G., Bairoch,A., Bridge,A., Bougueleret,L., Xenarios,I. and Viaria,A. (2012) UniPathway: a resource for the exploration and annotation of metabolic pathways. *Nucleic Acids Res.*, **40**, D761–D769.
18. Jain,E., Bairoch,A., Duvaud,S., Phan,I., Redaschi,N., Suzek,B.E., Martin,M.J., McGarvey,P. and Gasteiger,E. (2009) Infrastructure for the life sciences: design and implementation of the UniProt website. *BMC Bioinformatics*, **10**, 136.
19. Patient,S., Wieser,D., Kleen,M., Kretschmann,E., Martin,M.J. and Apweiler,R. (2008) UniProtJAPI: a remote API for accessing UniProt data. *Bioinformatics*, **24**, 1321–1322.
20. Kasprzyk,A. (2011) BioMart: driving a paradigm change in biological data management. *Database*, **2011**, bar049.
21. Vizcaino,J.A., Cote,R., Reisinger,F., Foster,J.M., Mueller,M., Rameseder,J., Hermjakob,H. and Martens,L. (2009) A guide to the proteomics identifications database proteomics data repository. *Proteomics*, **9**, 4276–4283.